International Journal of Recent Trends in Engineering & Research

ISSN (ONLINE) : 2455 – 1457

IMPACT FACTOR : 4.101

MULTI-MODAL RETRIEVAL IN NEWS FEED APP USING GCDL TECHNIQUE

Mr. J.Srinivasan[1],Mrs. J.Jayakrishnaveni[2]

[1]Assistant Professor, Department of Computer Science and Applications,
Adhiparasakthi College of Arts and Science,Kalavai
[2]Research Scholar, Department of Computer Science and Applications,
Adhiparasakthi College of Arts and Science,Kalavai

Abstract: Hashing methods have proven to be useful for a variety of tasks and have attracted extensive attention in recent years. Various hashing approaches have been proposed to capture similarities between textual, visual, and cross-media information. Nearest neighbor search methods based on hashing have attracted considerable attention for effective and efficient large-scale similarity search in computer vision and information retrieval community. Collective matrix factorization is a simple yet powerful approach to jointly factorize multiple matrices, each of which represents a relation between two entity types. we propose a novel method called semantic cross-media hashing (SCMH), which uses continuous word representations to capture the textual similarity at the semantic level and use a deep belief network (DBN) to construct the correlation between different modalities. To demonstrate the effectiveness of the proposed method, we evaluate the proposed method on three commonly used cross-media data sets are used in this work. Experimental results show that the proposed method achieves significantly better performance than state-of-the-art approaches. Moreover, the efficiency of the proposed method is comparable to or better than that of some other hashing methods.

I. INTRODUCTION

WITH the rapid expansion of theWorldWideWeb, digital information has become much easier to access, modify, and duplicate. Hence, hashing based similarity calculation or approximate nearest neighbour searching methods have been proposed and received considerable attention in recent years. Various applications such as information retrieval, near duplicate detection, and data mining are performed by hashing based methods. Due to the rapid expansion of mobile networks and social media sites, information input through multiple channels has also attracted increasing attention. Images and videos are associated with tags and captions. According to research published on eMarketer, about 75 percent of the content posted by Facebook users contains photos. The relevant data from different modalities usually have semantic correlations. Therefore, it is desirable to support the retrieval of information through different modalities. For example, images can be used to find semantically relevant textual information. On the other side, images without (or with little) textual descriptions are highly needed to be retrieved with textual query.

Along with the increasing requirements, in recent years, cross-media search tasks have received considerable attention. Since each modality having different representation methods and correlational structures, a variety of methods studied the problem from the aspect of learning correlations between different modalities. Existing methods proposed to use Canonical Correlation Analysis (CCA), manifolds learning, dual-wing harmoniums, deep autoencoder, and deep Boltzmann machine to approach the task. Due to the efficiency of hashing-based methods, there also exists a rich line of work focusing the problem of mapping multi-modal high-dimensional data to low-dimensional hash codes, such as Latent semantic sparse hashing (LSSH), discriminative coupled dictionary hashing (DCDH) , Cross-view Hashing (CVH), and so on.

Most of the existing works use a bag-of-words to model textual information. The semantic level similarities between words or documents are rarely considered. Let us consider the following examples:

S1. The company announces new operating system.
S2. The company releases new operating system.
S3. The company delays new operating system.

From these examples, we can observe that although only one word differs between the three sentences, sentence S3 should not be considered as the near duplicate sentence of S1 and sentence S2. The meaning expressed by S3 is much different with S1 and S2's. Since existing methods are usually based on lexical level similarities, this kind of issue cannot be well addressed by these methods.

In short text segments (e.g., microblogs, captions, and tags), the similarities between words are especially important for retrieval. For example: journey versus travel, coast versus shore. According to human-assigned similarity judgments, more than 90 percent of subjects thought that these pairs of words had similar meanings.

Motivated by the success of continuous space word representations (also called word embeddings) in a variety of tasks, in this work we propose to incorporate word embeddings to meet these challenges. Words in a text are embedded in a continuous space, which can be viewed as a Bag-of-Embedded-Words (BoEW). Since the number of words in a text is dynamic, we proposed a method to aggregate it into a fixed length Fisher Vector (FV), using a Fisher kernel framework . However, the proposed method only focus on textual information. Another challenge in this task is how to determine the correlation between multi-modal representations. Since we propose the use of a Fisher kernel framework to represent the textual information, we also use it to aggregate the SIFT descriptors of images. Through the Fisher kernel framework, both textual and visual information is mapped to points in the gradient space of a Riemannian manifold. However, the relationships that exist between FVs of different modalities are usually highly non-linear. Hence, to construct the correlation between textual and visual modalities, we introduce a DBN based method to model the mapping function, which is used to convert abstract representations of different modalities from one to another.

The main contributions of this work are summarized as follows.

- We propose to incorporate continuous word representations to handle semantic textual similarities and adopted for cross-media retrieval.
- Inspired by the advantages of DBN in handling highly non-linear relationships and noisy data, we introduce a novel DBN based method to construct the correlation between different modalities.
- A variety of experiments on three cross-media commonly used benchmarks demonstrate the effectiveness of the proposed method. The experimental results show that the proposed method can significantly outperform the state-of-the-art methods.

II. RELATED WORK

Along with the increasing requirement, extensive Hashing- based methods have been proposed for cross-media retrieval. In this section, we briefly describe the related works, which can be categorized into the following four research areas: cross-media retrieval, text Reuse detection, and hashing methods.

2.1 Cross-Media Retrieval

Cross-media retrieval, in which the modality of input query and the returned results can be of different, has received considerable attentions. We introduced a Canonical Correlation Analysis based method to construct isomorphic subspace and multi-modal correlations between media objects and polar coordinates to judge the general distance of media objects. Due to lack of sufficient training samples, relevance feedback of user was used to accurately refine cross-media similarities. Yang et al. proposed manifold-based method, in which they used Laplacian media object space to represent media object for each modality and an multimedia document semantic graph to learn the multimedia document semantic correlations. A rich-media object retrieval method is proposed to represent data consisting of multiple modalities, such as 2-D images, 3-D objects and audio files. To tackle the large scale problem, a multimedia indexing scheme was also adopted.

Since the relationships across different modalities are typically highly non-linear and observations are usually noisy, Srivastava and Salakhutdinov proposed a Deep Boltzmann Machine to learn joint representations of image and text inputs. The proposed model fuses multiple data modalities into a unified representation, which can be used for classification and retrieval. Xing et al. introduced to use dual-wing harmoniums to build a joint model for images and text. The model incorporated Gaussian hidden units together with Gaussian and Poisson visible units into a linear RBM model. In, a multimodal deep Boltzmann machine was proposed for learning multimodal data representations. To reduce the training time complexity,

2.2 Near-Duplicate Detection

The task of detecting near duplicate textual information has received considerable attentions in recent years. Previous works studied the problem from different aspects such as fingerprint extraction methods with or without linguistic knowledge, hash codes learning methods, different granularities, and so on.

Broder proposed Shingling method, which uses contiguous subsequences to represent documents. It does not rely on any linguistic knowledge. If sets of shingles extracted from different documents are appreciably overlap, these documents are considered exceedingly similar, which are usually measured by Jaccard similarity. In order to reduce the complexity of shingling, meta-sketches was proposed to handle the efficiency problem . In order to improve the robustness of shingle-like signatures, Theobald et al. introduced a method, SpotSigs. It provides more semantic pre-selection of shingles for extracting characteristic signatures from Web documents. SpotSigs combines stopword antecedents with short chains of adjacent content terms. The aim of it is to filter natural-language text pas-sages out of noisy Web page components. They also proposed several pruning conditions based on the upper bounds of Jaccard similarity.

I-Match is one of the methods using hash codes tore present input document. It filters the input document based on collection statistics and compute a single hash value for the remainder text. If the documents have same hash value, they are considered as duplicates. It hinges on the premise that removal of very infrequent terms and very common terms results good document representations for the near-duplicate detection task. Since I-Match signatures are respect to small modifications. The solution of several I-Match signatures, all derived from randomized versions in the original lexicon.

2.3 Hashing-Based Methods

In recent years, hashing-based methods, which create compact hash codes that preserve similarity, for single-modal or cross-modal retrieval on large-scale databases have attracted considerable attention . For single-modal, Hinton and Salakhutdinov proposed a two-layer network, which is called a Restricted Boltzmann machine (RBM), with a small central layer to convert high-dimensional input vectors into low-dimensional codes. In spectral hashing was defined to seek compact binary codes in order to preserve the semantic similarity between code words. The criterion used in spectral hashing is related to graph partitioning. Norouzi and Fleet introduced a method based on latent structural SVM framework for learning similarity preserving hash functions. A specific loss function is designed to incorporating both Hamming distance and binary quantization into consideration. In Self-Taught Hashing (STH) converted the hash codes learning problem into two stages. Unsupervised method, binarised Laplacian Eigenmap, is used to optimize l-bit binary codes. The classifiers were trained to predict the l-bit code for unseen documents.

A variety works studied the problem of mapping multimodal high-dimensional data to low-dimensional hash codes. Latent semantic sparse hashing proposed the use of Matrix Factorization to represent text and sparse coding to capture the salient structures of images. Then, these representations are mapped to a joint abstraction space. However, LSSH requires the use of both visual and textual information to construct the data set. Although out-of-samples can be estimated, the performances may be heavily influenced. Yu et al. introduced a discriminative coupled dictionary hashing approach, which generated a coupled dictionary for each modality based on category labels. Kumar and Udupa formulated formulated the problem of learning hash functions as a constrained minimization problem. Since the optimization problem is NP hard, they transformed it into a tractable eigenvalue problem by means of a relaxation. Inter-media hashing (IMH) uses a linear regression model to jointly learn a set of hashing functions for each individual media type Since we in this work learn the mapping functions between FVs of different modalities, all the hashing based methods for single modality can be incorporated into it.

2.4 Neural Networks for Representing Image and Text

The task of learning continuous space word representations have a long history as demonstrated outstanding results across a variety of tasks. Hinton and Salakhutdinov introduced a deep generative model to learn word-count vector and binary code for documents. In, the word representations are learned by a recurrent neural network language model. The proposed architecture consists of an input layer and a hidden layer with recurrent connections. Probabilistic neural network language model (NNLM) [48] simultaneously learns a distributed representation for each word and the probability function for word sequences. Bordes et al. proposed to use multi-task training process to jointly learn representations of words, entities and meaning representations. The work described in introduced a mix of unsupervised and supervised techniques to learn word vectors to capture both semantic and sentiment similarities among words.

On the image side, there are also a variety of studies tackling the problem of higher-level representations of visual information. Krizhevsky et al. proposed to use a deep convolutional neural network to perform object detection. In region proposals are combined with CNNs to generate features for object detection. Except for these supervised methods, unsupervised learning methods for training visual features have also been carefully studied. Lee et al. introduced convolutional deep belief network , a hierarchical generative model, represent images. Taylor et al. proposed a convolutional gated restricted Boltzmann machineto model the spatio-temporal features for videos.
Although, in this work, we use word embeddings and SIFT to represent texts and images respectively, the proposed method can also incorporate these representations.

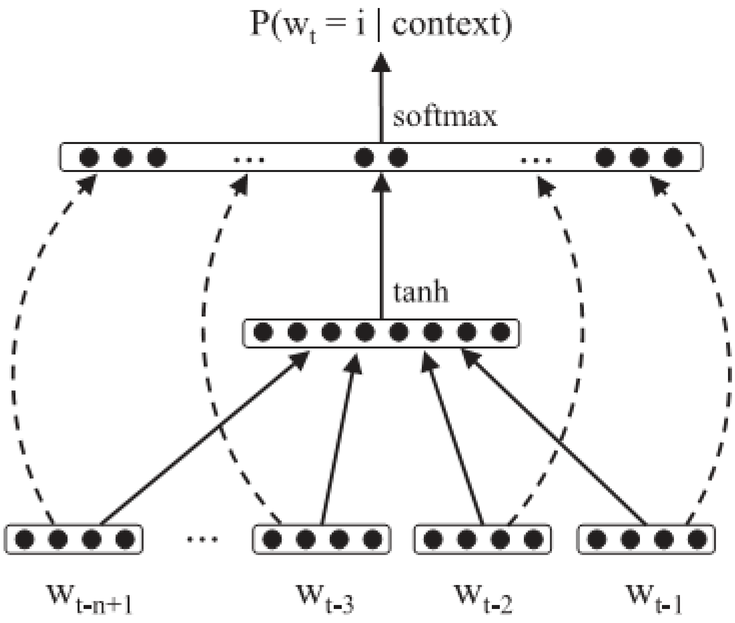
III. THE PROPOSED METHOD

The processing flow of the proposed semantic cross-media hashing (SCMH) method is illustrated in Fig. 2. Given a collection of text-image bi-modality data, we firstly represent image and text respectively. Through table lookup, all the words in a text are transformed to distributed vectors generated by the word embeddings learning methods. For representing images, we use SIFT detector to extract image keypoints. SIFT descriptor is used to calculate descriptors of the extracted keypoints. After these steps, a variable size set of points in the embeddings space represents the text, and a variable size set of points in SIFT descriptor space represents each image. Then, the Fisher kernel frame work is utilized to aggregate these points in different spaces into fixed length vectors, which can also be considered as points in the gradient space of the Riemannian manifold. Henceforth, texts and images are represented by vectors with fixed length. Finally, the mapping functions between textual and visual Fisher vectors (FVs) are learned by a deep neural network. We use the learned mapping function to convert FVs of one modality to another. Hash code generation methods are used to transfer FVs of different modalities to short length binary vectors. In the following section, we provide detailed examples of practical applications of the proposed method.

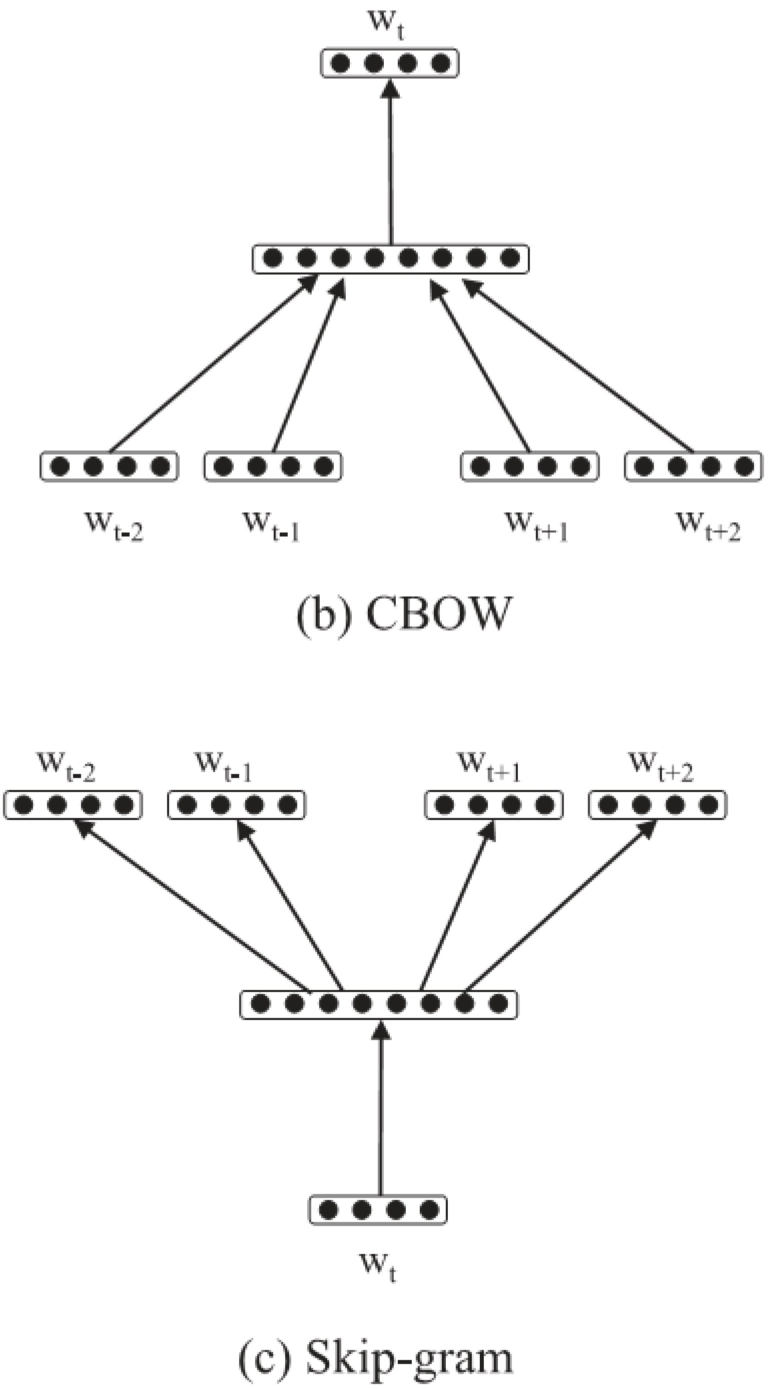
3.1 Word Embeddings Learning

Representation of words as continuous vectors recently has been shown to benefit performance for a variety of NLP and IR tasks. Similar words tend to be close to each other with the vector representation. Moreover, Mikolov et al. also demonstrated the learned word representations could capture meaningful syntactic and semantic regularities. Hence, in this work, we propose to use word embeddings to capture the semantic level similarities between short text segments.

Fig. 3 shows three architectures used for learning word embeddings. $w_i$represents the ith words in the given words sequence $\{w_1, w_2, \ldots, w_T\}$. Fig. 3a shows the architecture of the probabilistic neural network language model (NNLM) . It can have either one hidden layer beyond the word features mapping or direct connections from the word features to the output layer. Theyalso proposed to use function for the output layer to guarantee positive probabilities summing to 1. The word vectors and the parameters of that probability function can be learned simultaneously. In this work, we only use the learned word vectors.



(a) NNLM



(b) CBOW

(c) Skip-gram

Fig. 3. Methods used to learn word embeddings. The NNLM architecture predicates the probability of words based on the existing words. CBOW predicts the current word based on the context. Skip-gram predicts surrounding words given the current word

Figs. 3b and 3c show the architectures of the methods proposed by Mikolov. The architecture of CBOW, which is similar to NNLM, is shown in Fig. 3b. The main differences are that (i) the non-linear hidden layer is removed; (ii) the words from the future are included; (iii) the training criterion is to correctly classify the current ( ) word. The Skip-gram architecture, which is shown in Fig. 3c, is similar to CBOW. However, instead of predicting the current word based on the history and future words, it tries to maximize classification accuracy of words within a certain range before and after the current word based on only the current word as input.

Besides the methods mentioned above, there are also a large number of works addressing the task of learning distributed word representations. Most of them can also be used in this work. The proposed frame work has no limits in using which of the continuous word representation methods.

3.2 Fisher Kernel Framework

Fisher kernel framework was proposed to directly obtain the kernel function from a generative probability model. A parametric class of probability models $P(X|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^{l}$ for some positive integer l. If the dependence on θ is sufficiently smooth, the collection of models with parameters from Θ can be viewed as a manifold M_{Θ}. Though applying a scalar product at each point $P(X|\theta) \in M_{\Theta}$, it can be turned into a Riemannian manifold.

We denote a text or an image $X = \{x_i, 1 \leq i \, |X|$, where x_i is the embedding of th word of a text or the SIFT descriptors of the i th keypoint of an image, is the number of words in a text or the number of the extracted keypoints in an image. $|X|$ is D-dimension word embeddings or SIFT descriptors. We should note that there may be different parameters for different data sets. According to the Fisher kernel framework, X can be modeled by a probability density function. In this work, $P(X|\theta)$ is given by Gaussian mixture model (GMM), which a sum of N Gaussians $N(\mu_i, \sum_i)$ with weights w_i. Let $\theta = \{ w_i, \mu_i, \sum_i, \forall_i = 1 \dots N\}$ be the set of GMM parameters. The parameters θ are estimated through the optimization of Maximum Likelihood (ML) criterion using Expectation Maximization (EM) method.

Based on the learned parameters set θ, a text or an image X can be characterized by the gradient vector using the following function:

$$G_{\theta}^{X} = \nabla_{\theta} \log P(X|\theta)$$
$$= (\frac{\partial y}{\partial \theta_1} \log P(X|\theta), \dots, \frac{\partial y}{\partial \theta_l} \log P(X|\theta) \text{ ----(1)}$$

where G_{θ}^{X} is a vector whose dimensionality is only dependent on the number of parameters in λ , is only dependent on the number of words or keypoints. The gradient describes the contribution of each individual parameters to the generative process. It can also be interpreted as how these parameter contribute to the process of generating an example. We follow the work described in for normalizing these gradients by incorporating Fisher information matrix (FIM) F_{θ}. According to the theory of information geometry $\mathscr{u} = P(X|\theta), \theta \in \Theta$, which is a parametric family of distributions, can be regarded as a Riemanninan manifold Q with a local metric given by the FIM $F_{\theta} \in \mathbb{R}^{MXM}$:

$$F_{\theta} = E \, (\nabla_{\theta} \log P(X|\theta) \nabla_{\theta} \log P(X|\theta)^{T}) \text{-----(2)}$$

The similarity between two samples X and Y can be measured by the Fisher kernel defined as:

$$K_{Fk}(X,Y) = G_{\theta}^{X^T} F_{\theta}^{-1} G_{\theta}^{Y} \text{ ------------(3)}$$

Since F_{θ} is symmetric and positive definite, F_{θ}^{-1} can be transformed to $L_{\theta}^{T} \, L_{\theta}$ based on the Cholesky decomposition. Therefore, $K_{Fk}(X,Y)$ can be rewritten as follows:

$$K_{FK}(X,Y) = \mathscr{g}_{\theta}^{X^T} \mathscr{g}_{\theta}^{Y} \text{ ---------------(4)}$$

Where

$$\mathcal{g}_{\theta}^{X} = L_{\theta} G_{\theta}^{X} = L_{\theta} \nabla_{\theta} \log \mathrm{P}(\mathrm{X}|\theta) \text{ -------------(5)}$$

In this work, we assume that $x_i (1 \leq i \leq |X|)$ follows the naïve independence model, $\mathcal{g}_{\theta}^{X}$ can be rewritten as follows:

$$\sum_{i=1}^{|X|} L_{\theta} \nabla_{\theta} \log \mathrm{P}(\mathrm{x_i}|\theta) \text{ -------------(6)}$$

$\mathcal{g}_{\theta}^{X}$ is also referred to as the Fisher Vector of X.

3.3 Mapping Function Learning

To transfer the FVs of one modality to another, we propose to use a deep belief network with one hidden layer to achieve the task. Fig. 4 shows the structure of the proposed method. The building block of the network used in this work is the Gaussian restricted Boltzmann machine. Because we have converted both textual and visual information into the gradient space of a Riemannian manifold, we in this work use a single hidden layer model to do it.

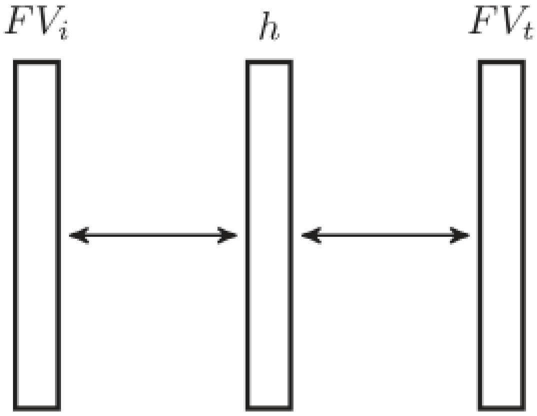


Fig. 4. A single hidden layer model for mapping FVs of different modalities. FV$_i$ and FV$_t$ denote the Fisher vector of image and text, respectively. h represents layer.

The restricted Boltzmann machine is a kind of an undirected graphical model with observed units and hidden units. The undirected graph of an RBM has an bipartite structure. It can be understood as a Markov random field with latent factors which explain the input observed data using binary hidden variables. Let v be the dimensional observed data, which can take real values or binary values. The dimension of stochastic binary units is . Each visible unit is connected to each hidden unit. The graphical model

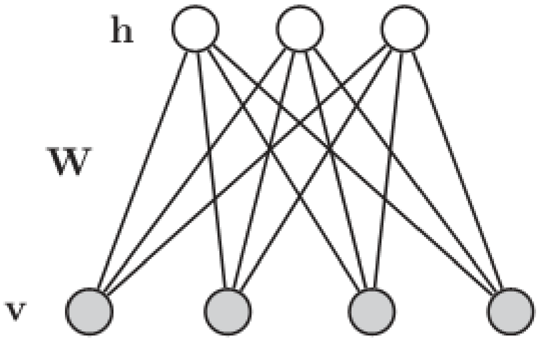


Fig 5. A graphical model representation of restricted Boltzmann machine.

representation is illustrated in Fig. 5. The parameters of RBM consist of the weight matrix $W \in \mathbb{R}^{LxK}$, the biases $c \in \mathbb{R}^{L}$for observed units, and the biases $b \in \mathbb{R}^{K}$for hidden units. If the observed units are real-valued, the model is called the Gaussian RBM. Its joint probability distribution can be defined as follows:

$$P(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{Z}\exp\big(-E(v,h)\big), E(v,h)$$

$$= \frac{1}{2\sigma^2}\sum_i (v_i - c_i)^2 - \frac{1}{\sigma}\sum_{i,j} v_i W_{ij} h_j - \sum_j b_j h_j \text{ --------(9)}$$

where Z is a normalization constant. The conditional distribution of this model can be written as follows:

$$P\left(h_j = 1\middle|V\right) = \ sigm\left(\frac{1}{\sigma}\sum_i W_{ij} h_j + b_j\right), \text{ ------(10)}$$

$$P(v_i = 1|h) = \mathcal{N}\left(v_i; \sigma\sum_j W_{ij} h_j + c_i, \sigma^2\right)\text{,----(11)}$$

Where $sigm(s) = \frac{1}{1+\exp(-s)}$ is the sigmoid function, and $N(.;.,.)$ is a Gaussian distribution. Although exact maximum likelihood learning in this model is intractable, sampling-based approximate maximum-likelihood methods can be used to estimated the parameters. Because the variables in a layer are conditionally independent, block Gibbs sampling can be performed in parallel. After training the RBM, Fisher vectors of different modalities can be transferred with the estimated parameters.

3.4 Hash Code Generation

Through the previous steps, a variable length of text segments or key points can be transferred to a fixed length vector. However, Fisher vectors are usually high dimensional and dense. It limits the usages of FVs for large-scale applications, where computational requirement should be studied. In this work, we propose to use hashing methods to address the efficiency problem.

The task of generating hash codes for samples can be formalized as learning a mapping $b(x)$, referred to as a hash function, which can project p-dimensional real-valued inputs $x \in \mathbb{R}^p$ onto q-dimensional binary codes $h \in \mathcal{H} \equiv \{-1,1\}^q$, while preserving similarities between samples in original spaces and transformed spaces. The mapping $b(x)$ can be parameterized by a real-valued vector w as :

$$b(x; w) = \ sign\left(f(x; w)\right)\text{, --------(12)}$$

Where sign(.) represents the element-wise sign function, and $f(x; w)$ denotes a real-valued transformation from $\mathbb{R}^p$to $\mathbb{R}^q$. In this work, Fisher vectors of text segments or keypoints are the x in mapping function $b(x; w)$. A variety of existing methods have been proposed to achieve this task under this framework using different forms of f and different optimization objectives. Most of the learning to hash methods for dense vectors can be used in this framework.

IV. EXPERIMENTS

To demonstrate the effectiveness of the propose method, we evaluated the following state-of-the-art methods on the three data sets.

- Cross-view Hashing maps similar objects to similar codes across the views to enable cross-view similarity search.

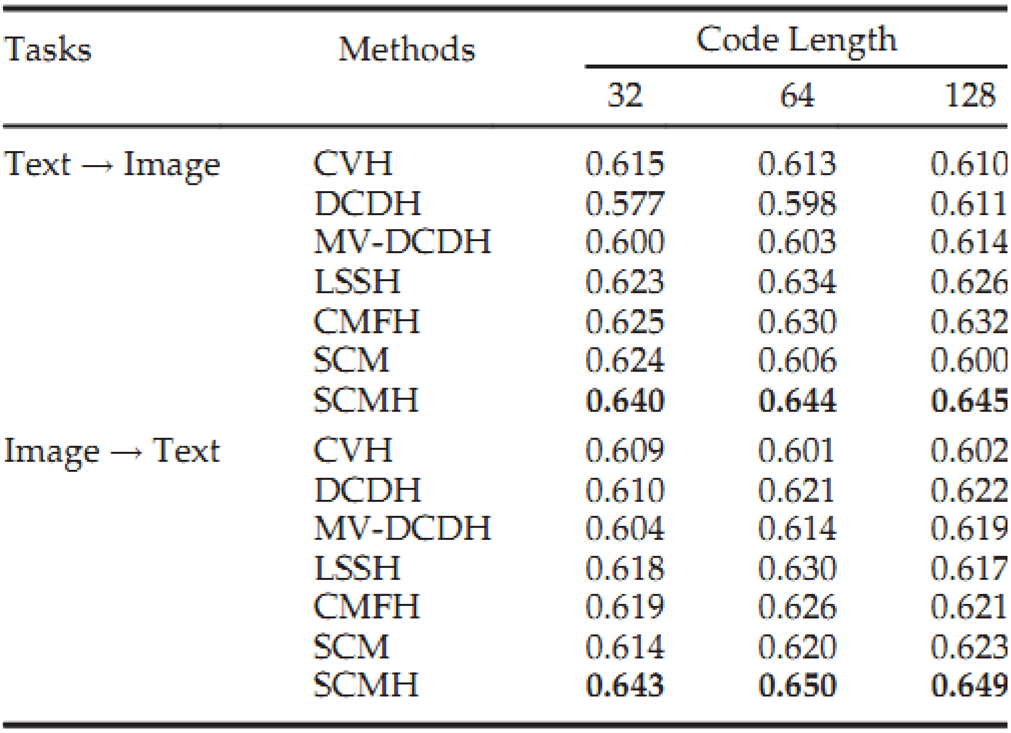
- Discriminative coupled dictionary hashing generates a coupled dictionary for each modality based on category labels.
- Multi view discriminative coupled dictionary hashing (MV-DCDH) is extended from DCDH with multi-view representation to enhance the represent ing capability of the relatively “weak” modalities.
- Latent semantic sparse hashing uses Matrix Factorization to represent text and sparse coding to capture the salient structures of images.
- Collective matrix factorization hashing (CMFH) generates unified hash codes for different modalities of one instance through collective matrix factorization with latent factor model.
- Semantic correlation maximization (SCM) integrates semantic labels into the hashing learning procedure for preserving the semantic similarity cross modalities.

The toolkits of LSSH, DCDH, and MV-DCDH are kindly provided by the authors. As we mentioned in the previous section, the proposed method SCMH can incorporate any hashing methods for single modality. In this work, we use Semantic Hashing to generate hash codes for both textual and visual information. Semantic Hashing is a multilayer neural network with a small central layer to convert high-dimensional input vectors into low-dimensional codes. For the length of hash codes, all the methods generate 32, 64, and 128 bits hash codes.

Following previous literatures on this task, we also adopt the widely used Mean Average Precision (MAP) as the evaluation metric. For a single query and top- retrieved instances, Average Precision (AP) is defined as follows:

$$AP = \frac{1}{R}\sum_{k=1}^{k} P(k)\delta(k),$$

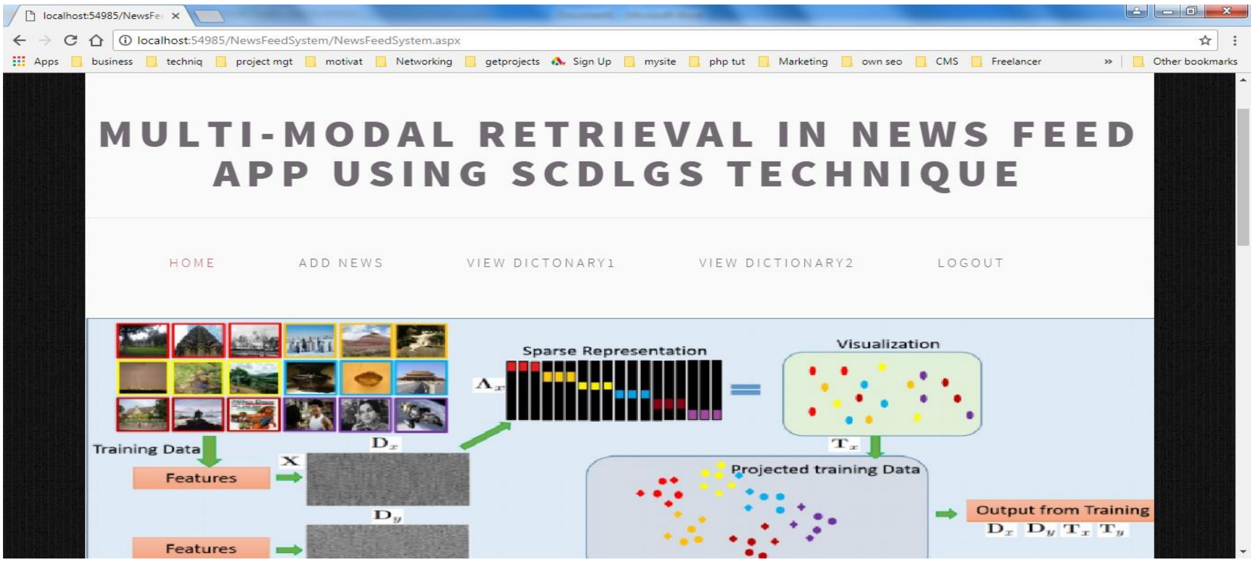
Where R denotes the number of ground-truth instances in the retrieved set, P(k) denotes the precision of top-k retrieved instances, and $\delta(k)$ is an indicator function which equals to 1 if the kth instance is relevant to query or 0 otherwise. In experiments. We set $K = 50$. Besides MAP, we also report precision-recall curve to represent the precision at different recall level.



Tasks	Methods	Code Length		
		32	64	128
Text → Image	CVH	0.615	0.613	0.610
	DCDH	0.577	0.598	0.611
	MV-DCDH	0.600	0.603	0.614
	LSSH	0.623	0.634	0.626
	CMFH	0.625	0.630	0.632
	SCM	0.624	0.606	0.600
	SCMH	**0.640**	**0.644**	**0.645**
Image → Text	CVH	0.609	0.601	0.602
	DCDH	0.610	0.621	0.622
	MV-DCDH	0.604	0.614	0.619
	LSSH	0.618	0.630	0.617
	CMFH	0.619	0.626	0.621
	SCM	0.614	0.620	0.623
	SCMH	**0.643**	**0.650**	**0.649**

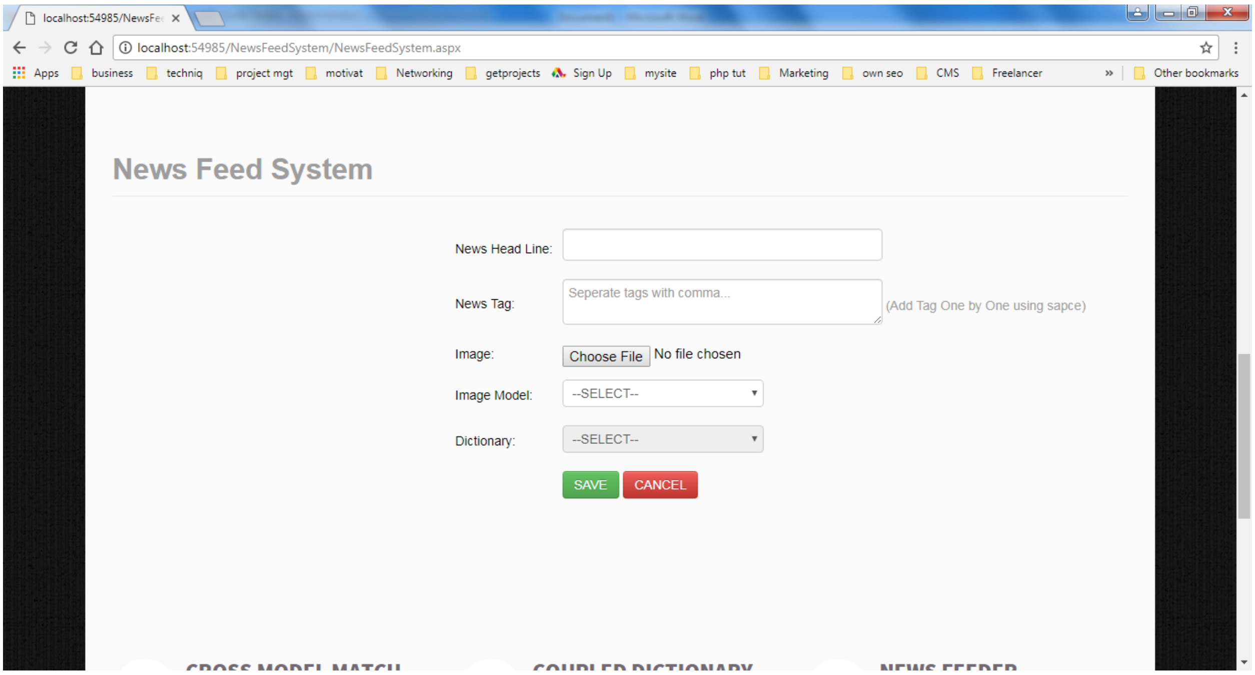
We report the results of Text→ Image and Image Text tasks on all three databases. For Text We report the results of Text Image and Image Text tasks on all three databases. For Text Image task, a text query, which contains the annotated tags of an image, is input to search images. The text query is firstly represented by a Fisher vector based on word embeddings. Then, the FV of text is mapped into a FV in image space. Finally, hamming distance is used to measure the similarities between the hash code of the converted FV and other hash codes of images. The top-*K* images are selected as the results. The procedure of Image →Text task is similar as the Text Image task. Since the Fisher vector mapping function needs training data, for each data set, we select 40 percent of the data to train the mapping function between text and image. 35 percent of the data are chosen as the retrieval database and the others are formed the query set. All the methods use the same data splits.

V. EVALUATION RESULT



localhost:54985/NewsFeedSystem/NewsFeedSystem.aspx

MULTI-MODAL RETRIEVAL IN NEWS FEED APP USING SCDLGS TECHNIQUE



localhost:54985/NewsFeedSystem/NewsFeedSystem.aspx

News Feed System

News Head Line:

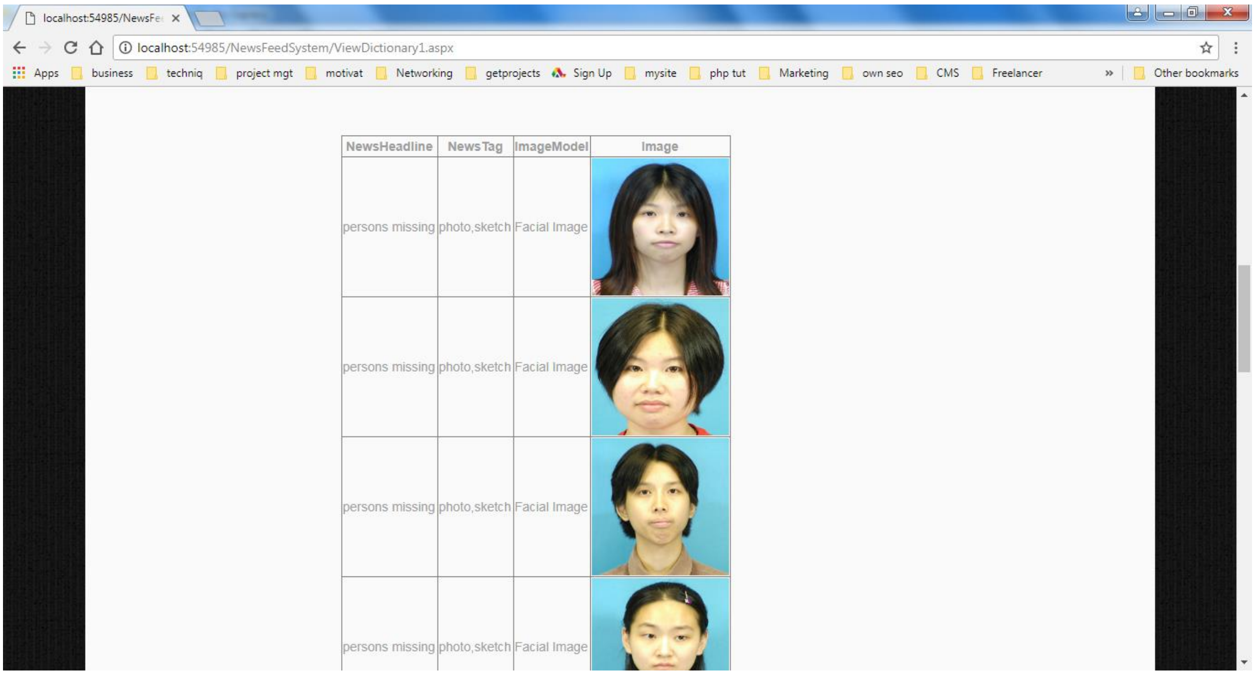
News Tag: Seperate tags with comma... (Add Tag One by One using sapce)

Image: Choose File No file chosen

Image Model: --SELECT--

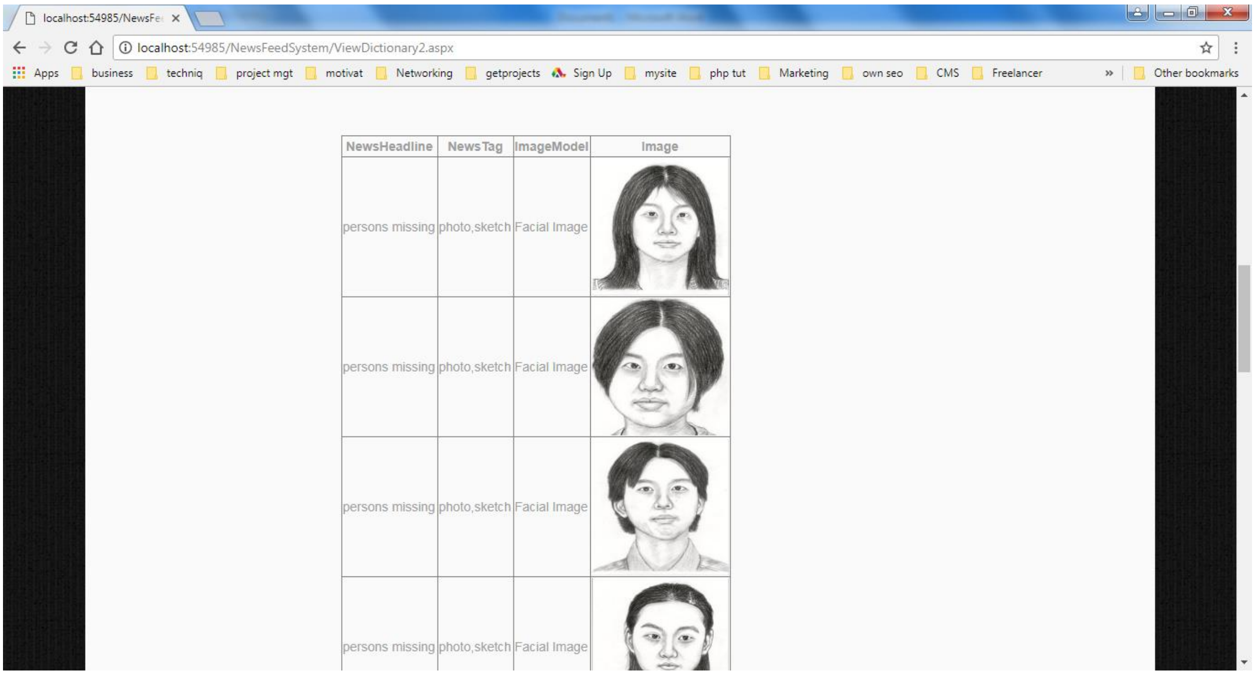
Dictionary: --SELECT--

SAVE CANCEL



localhost:54985/NewsFeedSystem/ViewDictionary1.aspx

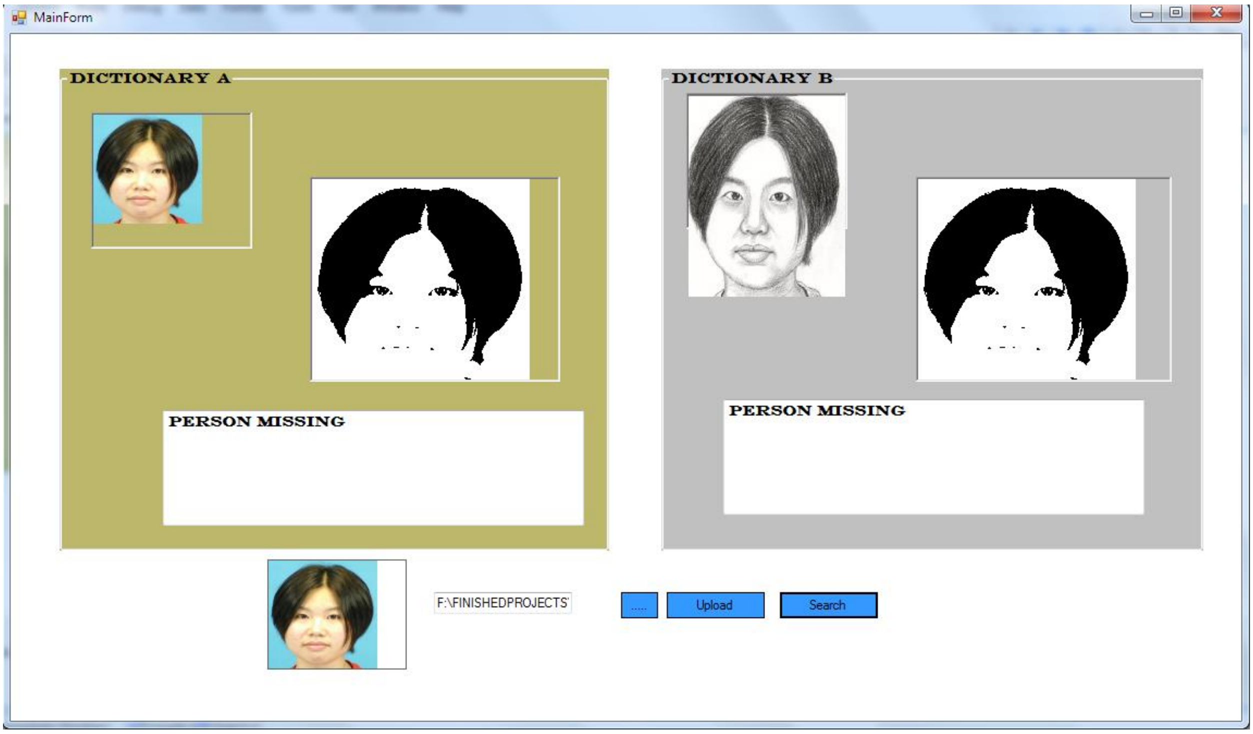
NewsHeadline	News Tag	ImageModel	Image
persons missing	photo,sketch	Facial Image	
persons missing	photo,sketch	Facial Image	
persons missing	photo,sketch	Facial Image	
persons missing	photo,sketch	Facial Image	



localhost:54985/NewsFeedSystem/ViewDictionary2.aspx

NewsHeadline	News Tag	ImageModel	Image
persons missing	photo,sketch	Facial Image	
persons missing	photo,sketch	Facial Image	
persons missing	photo,sketch	Facial Image	
persons missing	photo,sketch	Facial Image	

VI. SEARCH RESULT



VII. CONCLUSION

We propose a novel hashing method, SCMH, to perform the near-duplicate detection and cross media retrieval task. We propose to use a set of word embeddings to represent textual information. Fisher kernel framework is incorporated to represent both textual and visual information with fixed length vectors. For mapping the Fisher vectors of different modalities, a deep belief network is proposed to perform the task. We evaluate the proposed method SCMH on three commonly used data sets. SCMH achieves better results than state-of-the-art methods with different the lengths of hash codes. In NUS-WIDE data set, the relative improvements of SCMH over LSSH, which achieves the best results in these datasets, are 10.0 and 18.5 percent on the Text Image and Image Text tasks respectively. Experimental results demonstrate the effectiveness of the proposed method on the cross-media retrieval task.

VIII. FUTURE ENHANCEMENT

We can observe that the computational complexity of the proposed method is comparable with and state-of-the hashing methods. Comparing to the methods based on the matrix factorization, the proposed method is much more efficient. In future, we use semantic hash to generate hash codes of FVs. Hence, additional processing time is required to perform the calculation. However, if we use less complex hashing method, the efficiency can be further improved. It demonstrates that the proposed method is applicable for large scale applications.

REFERENCES

1. Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, “Click-through-based cross-view learning for image search,” in Proc. 37th Int. ACMSIGIR Conf. Res.Develop. Inf. Retrieval, 2014, pp. 717–726.
2. J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, “Inter-media hashing for large-scale retrieval from heterogeneous data sources,” in Proc. Int. Conf. Manage. Data, 2013, pp. 785–796.
3. D. Zhai, H. Chang, Y. Zhen, X. Liu, X. Chen, and W. Gao, “Parametric local multimodal hashing for cross-view similarity search,” in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 2754–2760.
4. D. Zhang and W.-J. Li, “Large-scale supervised multimodal hashing with semantic correlation maximization,” in Proc. 28th AAAI Conf. Artif. Intell., 2014, pp. 2177–2183

5. Y. Zhuang, Y. Yang, F. Wu, and Y. Pan, “Manifold learning based cross-media retrieval: A solution to media object complementary nature,” J. VLSI Signal Process. Syst. Signal, Image Video Technol., vol. 46, pp. 153–164, 2007.
6. E. P. Xing, R. Yan, and A. G. Hauptmann, “Mining associated text and images with dual-wing harmoniums,” in Proc. 21st Conf. Uncertainty Artif. Intell., 2005, pp. 633–641.
7. N. Srivastava and R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” in Proc. Adv. Neural Inf. Process. Syst.,2012, pp. 2222–2230.
8. S. Kumar and R. Udupa, “Learning hash functions for cross-view similarity search,” in Proc. Int. Joint Conf. Artif. Intell., 2011, pp. 1360–1365.
9. Q. Zhang, J. Kang, J. Qian, and X. Huang, “Continuous word embeddings for detecting local text reuses at the semantic level,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp. 797–806.
10. X. Wang, Y. Liu, D. Wang, and F. Wu, “Cross-media topic mining on wikipedia,” in Proc. 21st ACM Int. Conf. Multimedia, 2013, pp. 689–692.
11. H. Zhang, J. Yuan, X. Gao, and Z. Chen, “Boosting cross-media retrieval via visual-auditory feature analysis and relevance feedback,” in Proc. ACM Int. Conf.Multimedia, 2014, pp. 953–956.
12. M. Theobald, J. Siddharth, and A. Paepcke, “Spotsigs: Robust and efficient near duplicate detection in large web collections,” in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 563–570.